AP Stat Section 2.6 notes

***Analyzing 2 Categorical Variables***

* You can create categorical variables …

* Categorical variables are often shown …

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Frosh** | **Soph** | **Junior** | **Senior** | Total |
| **Male** |  |  |  |  |  |
| **Female** |  |  |  |  |  |
| Total |  |  |  |  |  |

Identify:

* Row variable
* Column variable
* Values of the variable
* Total (n)
* Cells
* Totals

***Example:*** Hospitals

|  |  |  |
| --- | --- | --- |
|  | **Hospital A** | **Hospital B** |
| **Died** | 63 | 16 |
| **Survived** | 2037 | 784 |

* What percent of people died?
* Of those people that went to Hospital A, what percent died?
* Of those people who went to Hospital B, what percent died?
* Of those people who died, what percent went to Hospital A?
* What percent of people died and went to Hospital B?

***2 types of Distributions for Categorical Variables***

1. **MARGINAL DISTRIBUTIONS**

* How to make:
* Looking for …
* ALWAYS …
* Example: Hair color vs. Gender

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Brown** | **Blonde** | **Black** | **Red** | **Total** |
| **MALE** | 26 | 24 | 10 | 3 | 63 |
| **FEMALE** | 20 | 35 | 12 | 6 | 73 |
| **TOTALs** | 46 | 59 | 22 | 9 | 136 |

* Find the marginal distribution for the HAIR COLOR variable
* Find the marginal distribution for the GENDER variable
* Represented Visually: BAR CHART

1. **CONDITIONAL DISTRIBUTIONS**

* Look at …
* Then look at …

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Brown** | **Blonde** | **Black** | **Red** | **Total** |
| **MALE** | 26 | 24 | 10 | 3 | 63 |
| **FEMALE** | 20 | 35 | 12 | 6 | 73 |
| **TOTALs** | 46 | 59 | 22 | 9 | 136 |

* Break down …
* ALWAYS …
* Example: Hair Color vs. Gender
* Find the conditional Distribution for the HAIR COLOR variable
* Find the conditional Distribution for the GENDER variable
* Represented visually: SEGMENTED (or STACKED) BAR GRAPH
  + Each bar = 100%
  + Values of variable on the x-axis
  + Bars are segmented into parts of each value

**AP Stat- worksheet 2.6A- Categorical Variables practice**

In a survey of adult Americans, people were asked to indicate their age and to categorize their political preference (liberal, moderate, conservative). The results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Liberal | Moderate | Conservative | *Total* |
| under 30 | 83 | 140 | 73 | *296* |
| 30 - 50 | 119 | 280 | 161 | *560* |
| over 50 | 88 | 284 | 214 | *586* |
| *total* | *290* | *704* | *448* | *1442* |

1. What proportion (percent) of the survey respondents were under the age of 30?
2. What proportion (percent) of the survey respondents were between 30 and 50 years old?
3. What proportion (percent) of the survey respondents were over the age of 50?
4. You have just calculated the ***marginal distribution*** for the ***AGE variable***. Make a bar graph of the marginal distribution for age.
5. Calculate the ***marginal distribution*** for the ***PREFERENCE variable*** (that is, calculate the percent of respondents for each of the 3 political preferences). Then make a bar graph of this marginal distribution.
6. Now let’s calculate the ***conditional distributions.*** Let’s start with the ***AGE variable.***
   1. First, list the values of the AGE variable
   2. Now focus on just the row for UNDER 30. Find the percent of JUST 30 YEAR OLDS that are liberal, moderate, and conservative. List them here.
   3. Now focus on just the row for 30 - 50. Find the percent of JUST 30 - 50 YEAR OLDS that are liberal, moderate, and conservative. List them here.
   4. Now focus on just the row for OVER 50. Find the percent of JUST OVER 50 YEAR OLDS that are liberal, moderate, and conservative. List them here.
   5. Now create a segmented (stacked) bar graph for the conditional distributions of AGE. Do this by putting the values of the AGE variable on the x-axis, then drawing a bar up to 100% for each value, and splitting each bar up into the percents for each of the political preferences.
7. Using your graph, write a few sentences commenting on whether there seems to be any relationship between age and political preference. In other words, does the distribution of political preference seem to differ among the 3 age groups? How so?
8. Now let’s calculate the ***conditional distributions*** for the ***POLITICAL PREFERENCE variable.***
   1. First, list the values of the POLITICAL PREFERENCE variable
   2. Now focus on just the column for LIBERAL. Find the percent of JUST LIBERALS that are 30, 30-50, and over 50. List them here.
   3. Now focus on just the column for MODERATE. Find the percent of JUST MODERATES that are 30, 30-50, and over 50. List them here.
   4. Now focus on just the column for CONSERVATIVE. Find the percent of JUST CONSERVATIVES that are30, 30-50, and over 50. List them here.
   5. Now create a segmented (stacked) bar graph for the conditional distributions of POLITICAL PREFERENCE below. LABEL THE AXES!!!

Go back to the original table for the following questions.

1. **Given that** someone is under 30, what is the chance (percent) that they are Liberal?
2. What percent of moderates were under 30?
3. What percent of the respondents were both over 50 ***and*** liberals?

**AP Stat- worksheet 2.6B- Categorical data examples**

1. A 4-year study reported in *The New York Times*, on men more than 70 years old analyzed blood cholesterol and noted how many men with different cholesterol levels suffered nonfatal or fatal heart attacks.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Low cholesterol | Medium cholesterol | High cholesterol |
| Nonfatal heart attacks | 29 | 17 | 18 |
| Fatal heart attacks | 19 | 20 | 9 |

1. Calculate the marginal distribution for cholesterol level (and make a bar graph). Do all work for each part of this question in the space provided below.
2. Calculate the marginal distribution for severity of heart attack (and make a bar graph).
3. Calculate three conditional distributions for the three levels of cholesterol (make a stacked bar graph).
4. Calculate the conditional distributions for the type of heart attack (make a stacked bar graph).

**AP Stat- Worksheet 2.6C- Independence**

1. A study was made to compare year in high school with preference for vanilla or chocolate ice cream with the following results.

|  |  |  |
| --- | --- | --- |
|  | **Vanilla** | **Chocolate** |
| Freshman | 20 | 10 |
| **Sophomore** | 24 | 12 |
| **Junior** | 18 | 9 |
| **Senior** | 22 | 11 |

1. Calculate the marginal distribution for year in school
2. Calculate the marginal distribution for flavor preference
3. Calculate the conditional distributions for each year in school. What do you notice? Where have you seen these numbers before?
4. Calculate the conditional distributions for of the two flavors. What do you notice? Where have you seen these numbers before?

**This is known as INDEPENDENCE.**

**INDEPENDENCE = When one variable does not affect the other variable**

**How do we tell if two variables are independent?**

* **If their marginal and conditional distributions match.**

1. An organization is concerned about the number of new employees that leave before they finish a year of work. So they predict whether the employee will stay by giving them a specific standardized test. Below are the results comparing the test prediction and the actual result.

|  |  |  |
| --- | --- | --- |
|  | **Actually stay** | **Actually leave** |
| **Predicted to stay** | 63 | 12 |
| **Predicted to leave** | 21 | 4 |

1. Of those employees that are predicted to stay, what proportion actually left?
2. Of those employees predicted to leave, what proportion actually left?
3. Is an employee that is predicted to stay any more likely to leave than an employee predicted to leave?
4. Find the marginal distribution for Prediction
5. Find the marginal distribution for Actual result
6. Find the conditional distribution for Prediction.
7. Find the conditional distribution for Actual result.
8. Are the two variables independent?

# AP Statistics- Worksheet 2.6D – Simpson’s Paradox

Suppose you need heart surgery and are trying to decide between two surgeons, Dr. Cardi and Dr. Kulp. Here are some data concerning their success rates:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Survived | Died | *Total* |
| Dr. Kulp | 800 | 200 | *1000* |
| Dr. Cardi | 900 | 100 | *1000* |
| *Total* | *1700* | *300* | *2000* |

1. Calculate the proportion of Kulp’s patients who survived and the proportion of Cardi’s patients who survived. Which doctor saved the higher percentage of patients?

Suppose we further categorize each patient according to whether they were in good or poor condition prior to treatment. The results are as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Good Condition | | |  | | Poor Condition | | | |  |
|  | | | Dr. Cardi | Dr. Kulp | |  | | Dr. Cardi | Dr. Kulp |
| Died | | | 590 | 10 | | Died | | 210 | 190 |
| Survived | | | 870 | 30 | | Survived | | 30 | 70 |

1. Verify that these two tables add together to form the table in the first question above.
2. Among those who were in good condition, compare the recovery rates for the two doctors. Which doctor saved the greater percentage of its patients who had been in good condition?
3. Among those who were in poor condition, compare the recovery rates for the two doctors. Which doctor saved the greater percentage of its patients who had been in poor condition?
4. Write a few sentences explaining how Dr. Cardi had the higher recovery rate overall, yet Kulp has the higher recovery rate for each type of patient. (hint: did one doctor tend to treat more of one type of patient?)
5. Which doctor would you rather see if you were ill? Why?

**This illustrates SIMPSON’S PARADOX.**